9-2-2005

Analysis of expressed sequence tags and the identification of associated short tandem repeats in switchgrass

Christian M. Tobias
USDA, ARS, christian.tobias@ars.usda.gov

Paul Twigg
University of Nebraska at Kearney, twiggp@unk.edu

Daniel M. Hayden
USDA, ARS

Kenneth P. Vogel
University of Nebraska-Lincoln, kvogel1@unl.edu

Robert B. Mitchell
University of Nebraska-Lincoln, rob.mitchell@ars.usda.gov

See next page for additional authors

Theor Appl Genet (2005) 111: 956–964
DOI 10.1007/s00122-005-0030-3

ORIGINAL PAPER

**Christian M. Tobias · Paul Twigg · Daniel M. Hayden
Kenneth P. Vogel · Rob M. Mitchell · Gerard R. Lazo
Elaine K. Chow · Gautam Sarath**

Analysis of expressed sequence tags and the identification of associated short tandem repeats in switchgrass

Received: 16 May 2005 / Accepted: 19 June 2005 / Published online: 28 July 2005

Abstract Switchgrass is a large, North American, perennial grass that is being evaluated as a potential energy crop. Expressed sequence tags (ESTs) were generated from four switchgrass cv. "Kanlow" cDNA libraries to create a gene inventory of 7,810 unique gene clusters from a total of 11,990 individual sequences. Blast similarity searches to SwissProt and GenBank non-redundant protein and nucleotide databases were performed and a total of 79% of these unique clusters were found to be similar to existing protein or nucleotide sequences. Tentative functional classification of 61% of the sequences was possible by association with appropriate gene ontology descriptors. Significant differential representation between genes in leaf, stem, crown, and callus libraries was observed for many highly expressed genes The unique gene clusters were screened for the presence of short tandem repeats for further development as microsatellite markers. A total of 334 gene clusters contained repeats representing 3.8% of the ESTs queried.

Communicated by T. Lübberstedt

C. M. Tobias (✉) · D. M. Hayden · G. R. Lazo · E. K. Chow
USDA, ARS, Western Regional Research Center,
Genomics and Gene Discovery Unit,
800 Buchanan Street, Albany, CA 94710, USA
E-mail: ctobias@pw.usda.gov
Tel.: +510-559-6172
Fax: +510-559-5818

P. Twigg
Biology Department, Bruner Hall of Science,
University of Nebraska at Kearney,
Kearney, NE 68849, USA

K. P. Vogel · R. M. Mitchell · G. Sarath
USDA, ARS, Wheat, Sorghum and Forage Research Unit,
Keim Hall, E.C. University of Nebraska,
Lincoln, NE 68583, USA

Introduction

Grasses belonging to the family Poaceae comprise, as a group, approximately 10,000 species and evolved about 55–70 million years ago (Kellogg 2001). The group is highly productive, covering approximately 20% of the earth's land surface and contains the world's grain crops, as well as important forage species. Several large perennial grasses such as switchgrass (*Panicum virgatum* L.) show potential for biomass production and utilization as energy crops.

Application of molecular approaches to plant-based biomass production in dedicated energy crops has only recently begun. Ultimately, these approaches may help improve yields, direct synthesis of valuable co-products, or after feedstock quality. Achievement of these goals would improve the economies of biomass production and conversion that have in the past limited its utilization. Switchgrass is a high yielding, C4, perennial grass component of the North American tall grass prairie that once occupied a large portion of the continental United States. It has been co-fired with coal in power plants but it also has excellent potential as a feedstock for ethanol production with yields estimated to be from 4,400 to 5600 l ha^{-1} (Moser and Vogel 1995; McLaughlin and Walsh 1998). An enhanced understanding of cell wall biogenesis at the molecular level in switchgrass may lead to new approaches to enhance feedstock quality for this end use.

Although genome size can vary greatly between grass species, the recent origin and diversification of grasses and their genetic similarity facilitates comparative genomic analysis (Devos and Gale 2000; Bennetzen and Ma 2003). Several other Panicoid grasses such as sugarcane, sorghum, and maize have been studied intensively and many of the molecular resources developed for these and other grasses may be transferable to switchgrass (Saha et al. 2004; Yu et al. 2004). The alignment of the switchgrass genome with that of other grasses could provide information on loci controlling

traits such as lignin and ash content that breeders could utilize. However, several features make switchgrass unique. It is extremely heteromorphic; upland and lowland ecotypes have been recognized, which can have variable numbers of chromosomes that range from $4x = 36$ to $8x = 72$ (Barnett and Carver 1967). It is cross-pollinated and displays a high degree of self-incompatibility. It is managed and has been selected for long-term persistence and productivity, as well as for forage quality.

Basic characterization of the switchgrass genome indicates that the tetraploid cultivars have a nuclear DNA content, approximately twice that of sorghum and about three and a half times that of rice (Hultquist et al. 1996). Chloroplast polymorphisms and random amplified polymorphic DNA (RAPD) markers have been used to evaluate diversity among cultivars and natural populations of switchgrass (Gunter et al. 1996; Hultquist et al. 1996). Addition of more useful nuclear markers will allow efficient indirect selection of traits which are too expensive or time consuming to measure directly.

As a starting point to facilitate certain biotechnological approaches, expressed sequence tag (EST) sequencing was chosen to rapidly survey switchgrass genes. ESTs are single pass sequences of randomly arrayed cDNA clones that provide enough sequence information (200–800 bp) to positively identify a given gene (Adams et al. 1991). Through comparison with molecular databases containing other annotated sequences and by analysing a sequence's differential representation in specific libraries, one can often make inferences about gene function. Very little sequence information from switchgrass has been deposited in public databases. Because EST projects can economically generate large gene inventories by sampling the expressed portion of the genome this strategy is useful for species such as switchgrass with sizeable genomes where direct sequencing would be wasteful. EST sequence information can also be mined for DNA sequence polymorphisms such as single nucleotide polymorphisms (SNPs) and microsatellites that can be used for genome characterization. We have generated 11,990 ESTs derived from four different cDNA libraries of the switchgrass cv. "Kanlow" a tetraploid, lowland cultivar, especially tolerant to flooding conditions.

Materials and methods

cDNA library construction

Panicum virgatum L. cv. "Kanlow" plants were raised from seed in a greenhouse at the University of Nebraska, Lincoln, under a 16 h ~26–30°C day/8 h ~ 22–26°C night growth regimen, using supplemental lighting from halide lamps (200 mol photons m^{-2} s^{-1}) in a soil mixture consisting of 40% Canadian peat, 40% coarse vermiculite, 15% masonry sand and 5% screened topsoil, amended with 7.5 lbs. Waukesha fine lime per cubic yard. Plants were watered biweekly with a nutrient solution containing 200 ppm N and with tap water as needed otherwise. Stem tissues were collected from mature flowering tillers (post-anthesis) based on the developmental index of Moore et al. 1991. The top three internodes below the peduncle were excised and the sheaths removed. Leaf tissue was collected by excising leaf blades selected at random from both mature and immature harvested tillers. Crown tissue was collected from six to eight week old plants that had started to tiller from which roots and tillers were removed. Callus was grown from mature caryopsis that had been surface sterilized and placed on callus initiation media with maltose as a carbon source (Somleva et al. 2002). Cultures were maintained in the dark at 28°C. The tissue contained a variety of callus morphologies. Plant and callus tissues were flash-frozen in liquid nitrogen and stored at −80°C until used for RNA isolation.

Total RNA was extracted from frozen stem internode, crown, leaf, and callus tissue using the Concert Plant RNA reagent (Invitrogen, Carlsbad, CA, USA). Messenger RNA was purified using the FastTrack 2.0 mRNA isolation system. First strand cDNA synthesis was primed with a *Not*I-oligo(T) adapter primer followed by second strand synthesis using the Superscript Plasmid System for cDNA Synthesis (Invitrogen). The resulting cDNA was ligated to *Sal*I adapters, digested with *Not*I, size selected, and cloned into the pSPORT1 cloning vector prior to transformation of Ultramax DH5αFT chemically competent *E. coli* (Invitrogen).

Automated DNA sequencing

The libraries were plated and individual colonies were robotically picked and arrayed into 384 well plates for long term storage. Sequencing templates were prepared using high-throughput methods adapted to 96-well microplates with a modified version of an alkaline lysis miniprep. Sequencing reactions were assembled using ABI PRISM BigDye terminator chemistry (Applied Biosystems, Foster City, CA, USA) on a PTC-225 thermocycler (MJ research, Hercules, CA, USA) with either M13 forward, T7, or M13 reverse sequencing primers. Excess dye-terminator nucleotides were removed by alcohol precipitation. The sequences were then subjected to electrophoresis on an ABI 3730 *xl* DNA sequencer (Applied Biosystems) using POP7 polymer. The resulting raw data files were processed using the Phred base-calling program (Ewing and Green 1998; Ewing et al. 1998). Phred also trimmed the sequences based on data quality using a probability cutoff value of 0.05 to retain only the high quality segment of the sequence. The trimmed sequences were further processed to mask the ends of reads that contained vector and adapter sequence using the program Cross-Match. Masked sequences were then removed from sequence and quality files using an in-house perl script (Lazo et al.

2004). Sequences less than 100 base pairs after processing were excluded from analysis. Sequences were then compared against selected databases containing vector, *E. coli*, bacteriophage λ, chloroplast, mitochondria, or rRNA sequences using the BlastN algorithm (Altschul et al. 1990) and those that aligned with an e-value $<10^{-20}$ were eliminated. Databases included: Univec (NCBI), *E. coli* (GenBank), mitochondria, plastid (GenBank) and rRNA (GenBank). All ESTs were then compared to SwissProt and GenBank non-redundant protein and nucleotide databases using either BlastN or BlastX and deposited in the dbEST division of GenBank using batch submission protocols from NCBI.

EST clustering

To produce a tentatively unique gene (TUG) set from the EST data, and obtain more protein coding region based on overlap of individual reads, the fasta sequence and corresponding quality files produced from Phred and subsequent processing steps were used as input for the Phrap assembly program. Phrap parameters (penalty −5; minmatch 50; minscore 100) resulted in EST clusters of >90% identity over a 100 bp window. This clustering process also served to assess the rate of new sequence discovery for each library, as sequencing progressed.

GO ontology

The BlastX algorithm was used to match ESTs to EBI UniProt Release 4.2 (Swiss-Prot Rel. 46.2, TrEMBL Rel. 29.2) (Apweiler et al. 2004). The UniProt gene associations to gene ontology (GO) terms were obtained from the Gene Ontology Consortium site (http://www.geneontology.org) using association tables that were provided. The available gene ontology database [March 2005] was also used as reference and data was uploaded and parsed from a MySQL database containing the information presented for GO. ESTs that were not annotated with GO terms were further matched to NCBI non-redundant database entries and tentative assignments made to descriptions to the NCBI Clusters of Orthologous Groups (COG) Index (http://www.ncbi.nlm.nih.gov/COG/). The entries that did not match this pool were further compared to NCBI dbEST entries.

Results

Sequencing summary

To produce a gene inventory for switchgrass, a total of 15,072 sequencing reads from four cDNA libraries were generated. These libraries were produced from stem internodes of flowering tillers (stem), leaf blades (leaf), crown tissue (crown), and callus tissue initiated from mature caryopses (callus). After processing to remove vector, low quality sequences, and those reads derived from contaminating sources such as rRNA, organelles, or bacterial contamination, a total of 11,990 reads of greater than 100 base pairs were deposited to the dbEST division of GenBank. Here, they received individual accession numbers DN140629-DN152624. Table 1 contains summary statistics for the project by library. The average length of high quality sequence submitted was 533 bp with average Phred score after trimming of 54.4; of these, 89% were read directionally from the 5′ end of the cDNA clone. Overall, the success rate for obtaining sequences of high quality from this project approached 80%. Sequences derived from the stem, callus, crown, and leaf cDNA libraries comprised, respectively, 35, 25, 20, and 20% of the collection.

EST clustering

After assembly of the individual sequences by Phrap, a total of 1,831 contigs and 5,979 singletons were obtained representing a total of 7,810 tentatively unique genes (TUGs). This number is likely to overestimate the number of genes actually sampled due to the possible presence of non-overlapping ESTs derived from the same gene, alternative splicing events, chimeric cDNA

Table 1 *Switchgrass cv. Kanlow EST summary*

Library	Reads	dbEST	5′	3′	Average Length	Average qual.	Source-specific ESTs (%)
Stem	5,167	4,168	4,106	62	567	52.4	58
Callus	3,744	3,019	3,019	0	516	54.9	62
Crown	3,552	2,462	2,083	379	496	54.0	69
Leaf	2,609	2,341	1,521	820	535	57.9	51
Total	15,072	11,990	10,729	1,261	533	54.4	

High quality sequences submitted to dbEST division of GenBank after filtering out rRNA, plastid, mitochodrial, and microbial contaminants. The average length of the ESTs, and Phred quality score for all the nucleotides in a specific library was determined after trimming. The source-specific ESTs within each library represent the number of contigs plus singletons derived from the specified library source after assembly by Phrap that are unique to that library expressed as a percentage of the total number of ESTs for each library

Table 2 *Diversity of EST sources in the TUGs*

TUGs	No. of libraries represented
7,172	1
552	2
84	3
2	4
7,810	Total

The number of libraries contributing ESTs to a TUG. For example, there are two TUGs that contain ESTs from all four libraries and 7172 TUGs that are represented by ESTs from a single library

clones, and allelic variants that could not be assembled by Phrap. The Phrap assemblies were conservative and did not join reads that differed by greater than 1–2%. After visual inspection of the contigs, over 90 single base substitutions were found in contigs of four or more ESTs that appear in at least two separate overlapping sequences. These represent potential SNPs based on the criteria of Picoult-Newberg et al. (1999), and could derive from intracultivar allelic variants at the same locus, homologous loci that show tetrasomic inheritance and/or paralogous loci within the genome.

For comparing the relative richness of gene diversity sampled from each library, library-specific contigs and singletons were compared. The crown library contained the greatest percentage of source-specific sequences with 69% of its assembled sequences found only in that library. By comparison, the leaf library was the least diverse with 51% of its assembled sequences being unique to that library. Table 2 shows the number of libraries represented by each TUG. Most TUGs are derived from a single library. These represent singletons and contigs containing multiple ESTs from the same library. Only two TUGs, one with similarity to a maize cysteine protease (Mir3) and the other with similarity to maize glutamine synthetase contain sequences from all four libraries.

Blast similarity searches to SwissProt, GenBank (release 144) non-redundant protein, and nucleotide databases were performed, and a total of 79% of the 7810 TUGs were similar to other protein or nucleotide sequences with an e-value threshold of $< 10^{-10}$ for BlastX. Of these, 27% were most similar to hypothetical or unknown proteins of plant origin.

Gene ontology

Of the 11,990 *P. virgatum* ESTs sequenced, 7,090 (59%) were able to be assigned by GO molecular terms, 6,867 (57%) were able to be assigned by GO Biological terms, and 6,058 (51%) were able to be assigned by GO Cellular terms. Of the remaining 4,225 ESTs, 2,950 had matches to the NCBI non-redundant database; a few of these sequences were able to be matched to the NCBI COG Index, but a majority of them (92% of the 2,950) were found to be unannotated genome derived sequences. Of those that did not match the NCBI non-redundant database, about 44% of the 1,275 sequences did match plant EST database entries. In summary of all 11,990 sequenced ESTs, 61% were placed in functional categories (GO plus COG annotated), 30% were unannotated and genome derived (from NR and dbEST), and 9% appeared to be not matching known sequences. Figure 1a shows the general range of sequence database matches and Fig. 1b shows the top GO annotations from the GO database matched group.

Library comparison

Potential differentially expressed sequences were identified based on library representation and these are shown in Table 3. Contigs containing five or more ESTs derived from a single library source were tallied by library, as were the number of ESTs represented by these contigs for purposes of comparison. Additionally, we compared the EST representation in the library of interest with the representation in all other libraries based on a transcriptional LOD score (Wu et al. 2004). A threshold LOD score of greater than three (less than 0.1% probability of the observed expression pattern arising purely by chance) was used for comparison. The leaf cDNA library yielded the greatest number of contigs in this category with 71 in spite of being represented by fewer sequences overall. These highly expressed sequences tended to be photosynthetic genes. The crown library contained the least number of differentially expressed contigs in this category containing three altogether.

Detailed pair-wise comparisons between the libraries are presented in Fig. 2. Six different pairings of the four sequenced libraries are possible. In each case, the overlap of two libraries was ascertained from the composition of the singletons and contigs assembled by Phrap and all classes of singletons and contigs were expressed as a percentage of the total of the two libraries. Contigs containing EST's from both libraries were further separated into three classes: (i) those that contained one EST from each library; (ii) those that contained one EST from one library and multiple ESTs from the second library; (iii) those that contained multiple ESTs from each library. The crown and stem libraries overlapped to the greatest degree. Most of this overlap is due to contigs containing one EST from each library, as is the case with the other library comparisons. Comparison of leaf library ESTs with the remaining three libraries highlights the lack of substantial overlap between them. An exceeding small number of ESTs from the leaf library formed contigs with the other libraries (1.2% with stem and 0.7% with crown). This is in part due to the significantly lower gene diversity (Table 1) present in the leaf library and in part due to fewer sequenced ESTs as a whole from that library.

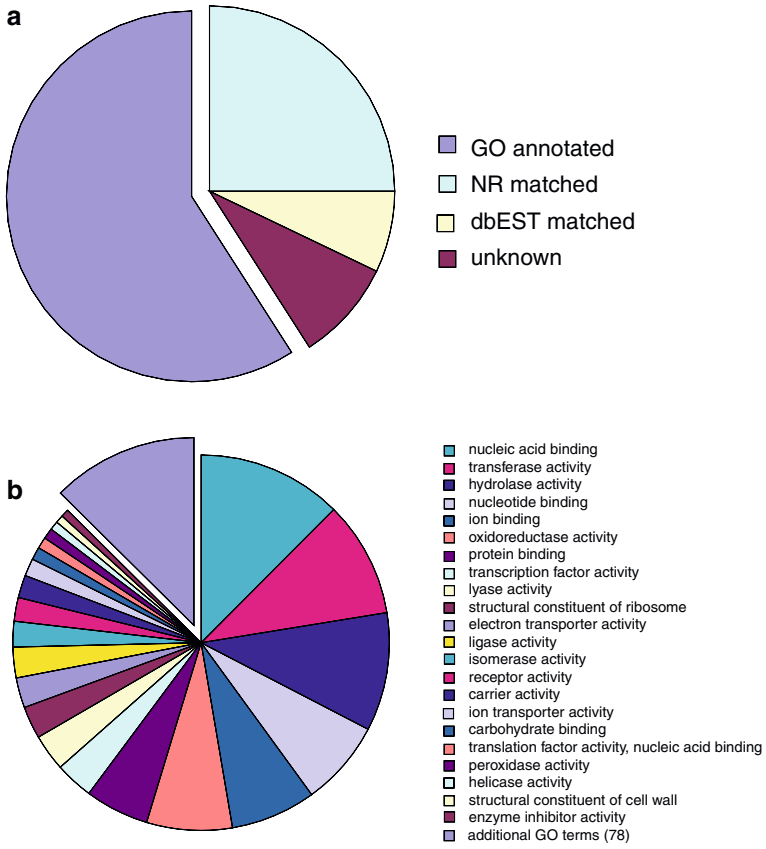

Fig. 1 Sequence annotation and functional classification. (**a**) To build EST sequence annotations, *P. virgatum* ESTs were first compared to the UniProt database and GO assignments determined, and remaining sequences compared to NCBI non-redundant database and then to the NCBI dbEST database. Shown are the breakdowns for sequence matches. (**b**) The proportional distribution of the top GO molecular assignments (at level 2) are shown and listed; the last category is a combination of 78 other GO classifications

Table 3 *Differential* Representation of ESTs Among Libraries

Source (library)	No. of Contigs	No. of ESTs	Contigs with LOD > 3	Example of gene products
Crown	8	43	3	Metallothionein-like protein (6 ESTs), Root specific lectin precurser (6 ESTs)
Leaf	71	593	71	Photosystem I F-subunit precurser (27 ESTs), Photosystem I reaction center subunit II/PSI-D (20 ESTs)
Stem	61	451	32	Chl a + b binding prot. precurser (14 ESTs), BTH-induced ERF transcriptional factor 4 (11 ESTs)
Callus	26	147	10	Fructose bisphosphate aldolase (9 ESTs), Alpha-amylase/subtilisin inhibitor (7 ESTs), Alcohol dehydrogenase (6 ESTs)

ESTs derived from a single library that Phrap assembled into contigs containing at least five members were tallied and assigned putative identifications based on top scoring Blast hit. These represent abundant transcripts likely to be preferentially expressed in the tissues used to prepare individual libraries. The total number of ESTs represented by the contigs is reported. The number of these contigs with transcriptional LOD scores (Wu et al. 2004) above three are shown

Representation of Genes involved in Cell Wall Biogenesis in Switchgrass EST Collection

Since cell wall biogenesis represents a particularly important target for manipulation in switchgrass, we focused on this area for more detailed analysis of gene representation. Gene families that play roles in cell-wall biogenesis have been broadly categorized along functional lines which include: (i) generation of cell wall precursors; (ii) polysaccharide synthases and glycosyl transferases; (iii) secretion and targeting pathways; (iv) assembly, architecture, and growth; (v) differentiation and secondary cell wall formation; and (vi) signaling and response mechanisms. Blast annotations from the best matches to existing protein and nucleotide databases, as well as individual Blast results against specific cell wall proteins, were analyzed to determine the representation of these sequences in the switchgrass EST collection and the results are shown in Table 4. Most gene categories and subcategories listed were represented by at least two individual ESTs. Categories that were poorly represented or missing include callose synthase, xyloglucan fucosyltransferase, glycoside hydrolase family 17, pectin/pectate hydrolases and lyases, rhamnogalacturonan I

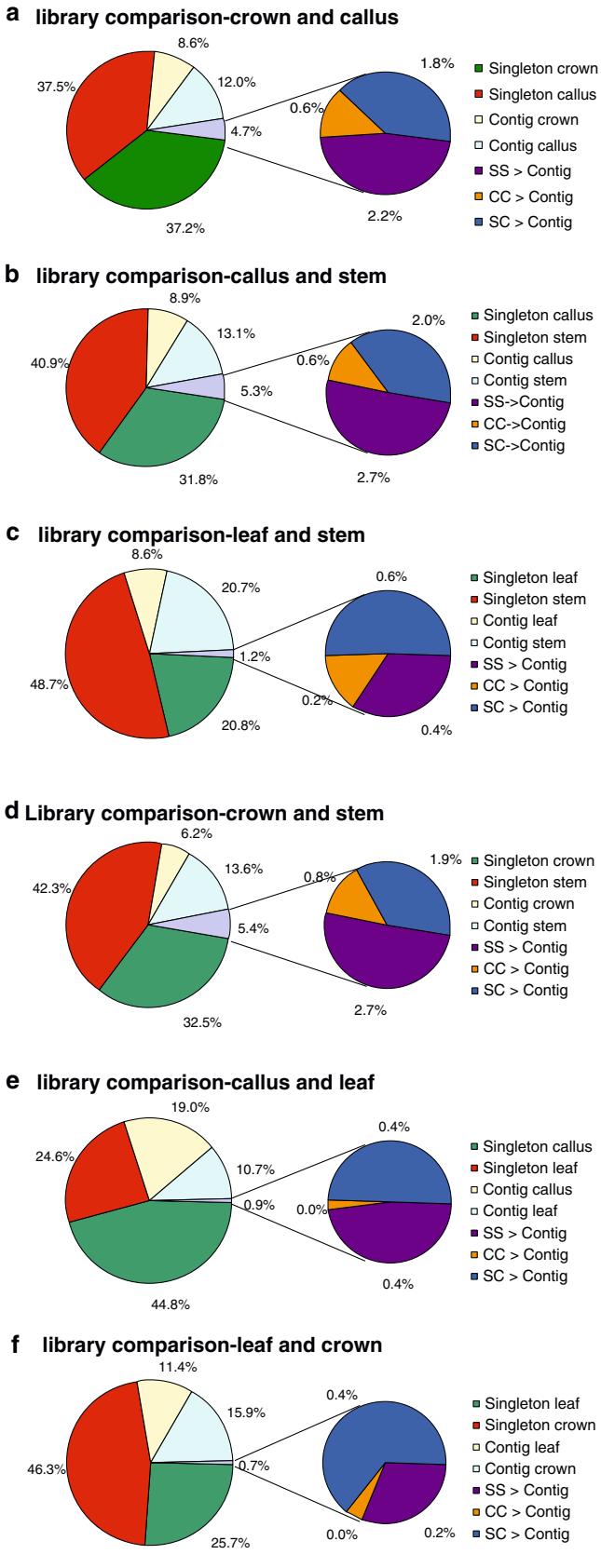

Fig. 2 Overlap of Switchgrass libraries. The ESTs derived from each library were assembled using the Phrap program (Ewing and Green 1998; Ewing et al. 1998). See Methods for assembly conditions. ESTs for which there were no matches or that could not be assembled into contigs were considered singletons. Pairwise comparisons between singletons and contigs derived from each library are shown: (**a**) crown and callus library overlap, (**b**) callus and stem library overlap, (**c**) leaf and stem library overlap, (**d**) crown and stem library overlap,(**e**) callus and leaf library overlap, and (**f**) leaf and crown library overlap. The percentages of each class are indicated by the legend are shown next to the colored wedges for the two libraries being compared. Proceeding in a counter clockwise direction from the top of the *left circle*, the classes are: singletons present in library A (*red*), singletons present in second library B (*green*), overlap between the two libraries (*light purple*), contigs containing ESTs from A (*light blue*), and contigs containing ESTs from library B (*light yellow*). The overlapping portion of the libraries comprising the light purple wedge is further represented in the smaller chart on the right. *CC* > Contig, multiple ESTs from each library (*orange*); *SS* > Contig, single ESTs from each library (*purple*); *SC* > Contig, single EST from one library and multiple ESTs from the other (*blue*)

lyases, feruloyl esterases, and pectin acetylesterases. The most abundant categories comprised genes with putative roles in lignin precurser biosynthesis and polymerization such as phenylalanine ammonium lyase, o-methyltransferases, CoA-ligases, laccases, and peroxidases.

Sequence polymorphism

To obtain gene-derived markers for further genetic studies, the contig assembly was screened for potentially polymorphic microsatellite sequences consisting of stretches of di-, tri-, and tetrameric nucleotide repeats by running the perl script SSRIT that was developed for this purpose (Temnykh et al. 2001). We used parameters that would detect dimeric motifs with nine or more repeats, trimeric motifs with six or more repeats and tetrameric motifs with five or more repeats as these have been associated with higher degrees of polymorphism in general (Cho et al. 2000; Weber 1990). A total of 334 potential microsatellite loci were returned representing 3.8% of the total number of ESTs queried. This is similar to the rate of microsatellite discovery in wheat, rice, maize, sorghum, and barley ESTs. The relative frequency of di-, tri-, and tetranucleotide repeats is also similar to that found in EST collections of other grasses (Kantety et al. 2002). The different classes of repeats are shown in Table 5. Trinucleotide repeats represented 79% of the total number of microsatellites. Of these, (CCG)*n* repeats were the most common. Dinucleotide repeats were also detected with (GA)*n* and (AT)*n* repeats being the most common classes. No (GC)*n* repeats were returned, and tetranucleotide repeats were also relatively uncommon. The simple sequence repeat (SSR) containing contig

Table 4 *Representation of Cell Wall Gene Families in Switchgrass ESTs*

Gene Family	Library				
	Callus	Crown	Leaf	Stem	Total
1. Pathways of substrate generation					
1.1 Nucleotide-sugar interconversion pathways	5	8	0	17	30
1.2 C-1 kinases and sugar salvage pathways	4	3	1	2	10
1.3 Pathways of phenylpropanoid biosynthesis	4	19	11	41	75
2. Polysaccharide synthases and glycosyl transferases					
2.1 Cellulose synthases	2	2	1	2	7
2.2 Cellulose synthase-like Genes	0	3	0	1	4
2.3 Glycosyl transferases	0	2	6	7	15
2.3.1 GT family 8	0	1	0	1	2
2.3.2 GT family 47	1	1	0	1	3
2.3.3 Xyloglucan Fucosyltransferase					0
2.4 Callose synthase genes					0
3. Secretion and targeting pathways					
3.1 Vesicle trafficking	2	0	0	1	3
3.2 Cytoskeleton-associated proteins	3	0	1	8	12
3.3 Plasma membrane fusion					ND
3.4 Endocytosis					ND
3.5 Cell-plate formation					ND
4. Assembly, architecture, and growth					
4.1 Growth modifying proteins					
4.1.1 Expansins	2	3	0	4	9
4.1.2 Yieldins	4	1	0	0	5
4.2 Xyloglucan endotransglucosylase/hydrolases	0	2	4	9	15
4.3 Hydrolases	0	0	0	7	7
4.3.1 Exo-acting glycanases	0	0	0	2	2
4.3.1.1 ß-Galactosidase family 35	3	0	0	2	5
4.3.2 Endo-acting glycanases					
4.3.2.1 Glycoside hydrolase family 9	5	3	0	6	14
4.3.2.2 Glycoside hydrolase family 17	0	0	0	1	1
4.3.3 Pectin/Pectate hydrolases	0	0	0	1	1
4.4 Lyases					
4.4.1 Pectate and pectin Lyases	0	1	0	0	1
4.4.2 Rhamnogalacturonan I lyases	0	0	0	0	0
4.5 Esterases					
4.5.1 Pectin methyl esterases	0	2	3	2	7
4.5.2 Pectin acetylestereases	0	0	0	0	0
4.5.3 Feruloyl esterases	0	0	0	0	0
4.6 Structural proteins					
4.6.1 Hydroxyproline-rich glycoproteins	0	2	0	3	5
4.6.2 Proline-rich proteins	1	1	0	5	7
4.6.3 Glycine-rich proteins	3	4	2	1	10
4.6.4 Arabinogalactan-proteins	0	1	0	8	9
5. Differentiation and secondary wall formation					
5.1 Lignan synthesis	0	1	0	1	2
5.2 Lignin assembly and modification					
5.2.1 Laccases	0	2	0	5	7
5.2.2 Peroxidases	12	6	1	5	24
5.2.3 Germin-like proteins	2	6	1	7	16
6. Signaling and response mechanisms					
6.1 Generation of signal molecules					ND
6.2 Reactive-oxygen species generation	0	0	0	1	1
6.3 Receptor-like kinases and their ligands	9	6	5	4	24
6.4 GPI-anchored proteins	1	0	0	3	4
Total	63	80	36	158	337

Top cell wall related Blast hits to ESTs were tabulated and categorized with respect to gene families believed to be involved in cell wall biogenesis (http://cellwall.genomics.purdue.edu/)

assemblies were compared using BlastN to the consensus sequences of 2,463 nonredundant Uni-EST-SSR contigs derived from an assembly of SSR-containing wheat, rice, maize, sorghum, and barley genes (Yu et al. 2004). These represent conserved microsatellite containing genes that could be used for cross-species comparisons. A total of 68 switchgrass microsatellites produced significant (e-value $< 10^{-10}$) alignments with this set of conserved grass microsatellites. A set of 96 primers pairs was synthesized to the switchgrass microsatellites. Of these, 46 primer sets amplified well. These were used for screening polymorphism between several individuals of switchgrass cvs. ''Kanlow'' and ''Alamo''. At least one polymorphic band was detectable for each primer tested with an average of 2.72 amplicons per genotype (unpublished data).

Discussion

There is a growing need for renewable sources of energy as alternatives to fossil fuels. As the energy within biomass is ultimately derived from the sun and is very abundant, this will be an increasingly important, greenhouse-gas neutral source of energy. Switchgrass represents one of the several high yielding herbaceous species that is being considered as a dedicated energy crop. Its broad genetic base provides the raw diversity necessary for further improvements through artificial selection, and its high yields, wide adaptability, and relatively low production costs are several reasons that the Department of Energy's Feedstock Development Program chose switchgrass for part of its research on biomass to ethanol conversion.

Relative to other EST sequencing efforts, the total number of switchgrass ESTs generated in these work places it eighth among the grass species listed in the dbEST database although several species with ongoing EST projects are not listed. The large majority of the switchgrass sequences were derived from 5′ end sequencing to obtain as much coding sequence as possible and avoid sequencing through the poly(A) tail, which can result in poor sequence quality.

Comparisons of different libraries involved in the sequencing showed the most library overlap (5.4%) between stem and crown libraries. This overlap was not unexpected because the crown is actually compressed stem tissue in switchgrass. However, this conclusion is somewhat contradicted by the presence in the crown library of a highly represented sequence with similarity to a root-specific barley lectin (Lerner and Raikhel 1989). Without root library sequences for comparison, there are several possible interpretations of this data. In contrast to other sources of ESTs, the leaf library derived ESTs overlapped very little with other libraries and contained the least gene diversity of three libraries (Table 1) due to the large representation of leaf-specific genes including many highly expressed genes with unambiguous roles in photosynthesis.

Analysis of cell-wall related gene families present in the switchgrass EST collection focused on the major processes described at the cell-wall genomics web site (cellwall.genomics.purdue.edu, Purdue University). Six broad functional processes that pertain to wall functions are defined, and we found switchgrass ESTs related to these processes to be well represented in most cases. Exceptions included genes involved in pectin biosynthesis and degradation, which were underrepresented. Other missing categories include xyloglucan fucosyltransferase and callose synthase. The type II cell walls present in grasses are pectin poor, and they also do not incorporate fucose into their xyloglucans (Carpita 1996) thus these findings are in agreement with expectations. Callose, a β1–3 glucan is produced in grasses and putative callose synthase genes are present in both rice and *Arabidopsis*. Therefore, their absence in the switchgrass EST collection was unexpected and may be due to low sequence coverage. In contrast, genes involved in secondary cell wall biogenesis were well represented, particularly in the stem cDNA library. Stem tissue is highly lignified so it is not surprising that due to the large number of stem ESTs generated, these represent the largest class of sequences, which includes genes with roles in monolignol biosynthesis and lignin polymerization. Altering expression of many of these genes has been shown to alter lignin content and/or composition (Anterola and Lewis 2002) that can result in improved digestibility or pulping properties. Therefore, this class of genes in switchgrass could be used for transgenic approaches to alter feedstock quality in a manner that might allow more economical conversion to simple sugars, or conversely greater energy per unit mass due to the high heat value of lignin. Although transformation technology exists for this species (Richards et al. 2001; Somleva et al. 2002), it has not yet been used for crop improvement.

Mining of EST sequence data for the presence of microsatellites can be a productive way of obtaining gene-associated markers. The most abundant classes of microsatellites in the switchgrass EST collection are the GC-rich trinucleotide repeats that also are the most abundant in human and maize cDNAs (Jurka and Pethiyagoda 1995; Chin et al. 1996). However, these classes exhibit less sequence length variability than dinucleotide repeats (Chakraborty et al. 1997; Cho et al. 2000). The highly variable (GA)*n* repeats that are most efficiently amplified in rice (Temnykh et al. 2001), were the largest class of dinucleotide repeat in the switchgrass. Evolutionary constraints acting on EST-derived microsatellites may limit sequence variability; however, in naturally outcrossing species, EST-derived microsatellites are in general more variable than in self-pollinating species. For white clover and tall fescue, 67 and 66% respectively of EST-derived microsatellites detected

Table 5 *Frequency of short tandem repeats in EST collection*

Class	Repeat	Frequency
Di-	(GA)*n*	37
	(AC)*n*	4
	(GC)*n*	0
	(AT)*n*	13
Tri-	(CCG)*n*	144
	(ACG)*n*	14
	(AGG)*n*	21
	(ACC)*n*	22
	(AAG)*n*	9
	(ACT)*n*	1
	(AAC)*n*	5
	(AAT)*n*	2
	(AGC)*n*	43
	(ATC)*n*	1
	(ATG)*n*	2
Tetra-	All	16

Assembled contigs were screened with SSRIT to identify repeats equal to or greater than 18 bp

some polymorphism compared to 43% for rice (Barrett et al. 2004; Saha et al. 2004).

The EST sequence resources for switchgrass we have generated should provide readily available sources of genes for further research and sequence data that can be used to discover and develop gene-linked markers such as microsatellites or SNPs for marker assisted breeding efforts, identity-preservation, and diversity assessment in this species or in closely related species such as proso millet. Traditional breeding efforts have resulted in the release of improved cultivars with increased digestibility (Vogel et al. 1991; Vogel et al. 1996). These efforts should be augmented by the availability of new molecular resources.

Acknowledgments Supported by the United States Department of Agriculture, Agricultural Research Service CRIS 5325-21000-013-00, NP307 Biofuel and Bioenergy Alternatives. This work was also supported in part by NIH Grant P20 RR16569 from the BRIN Program of the National Center for Research Resources, and by a University of Nebraska at Kearney Research Services Council grant.

References

Adams M, Kelley J, Gocayne J, Dubnick M, Polymeropoulos M, Xiao H, Merril C, Wu A, Olde B, Moreno R et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Altschul SF, Gish W, W M, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Anterola AM, Lewis NG (2002) Trends in lignin modification: a comprehensive analysis of the effects of genetic manipulations/mutations on lignification and vascular integrity. Phytochem 61:221–294

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) UniProt: the universal protein knowledge base. Nucleic Acids Res 32:D115–D119

Barnett FL, Carver RF (1967) Meiosis and pollen stainability in switchgrass, *Panicum virgatum* L. Crop Sci 7:301–304

Barrett B, Griffiths A, Schreiber M, Ellison N, Mercer C, Bouton J, Ong B, Forster J, Sawbridge T, Spangenberg G, Bryan G, Woodfield D (2004) A microsatellite map of white clover. Theor Appl Genet 109:596–608

Bennetzen J, Ma J (2003) The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. Curr Opin Plant Biol 6:128–133

Carpita NC (1996) Structure and biogenesis of the cell walls of grasses. Ann Rev of Plant Physiol Plant Mol Biol 47:445–476

Chakraborty R, Kimmel M, Strivers D, Davison L, Deka R (1997) Relative mutation rates at di- tri- and tetranucleotide microsatellite loci. Proc Natl Acad Sci USA 94:1041–1046

Chin E, Senior M, Shu H, JSC S (1996) Maize simple repetitive DNA sequences: abundance and allele variation. Genome 39:866–873

Cho Y, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch S, Park W, Ayres N, Cartinhour S (2000) Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). Theor Appl Genet 100:713–722

Devos K, Gale M (2000) Genome relationships: the grass model in current research. Plant Cell 12:637–646

Ewing B, Green P (1998) Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186–194

Ewing B, Hillier L, Wendl M, Green P (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185

Gunter LE, Tuskan GA, Wullschleger SD (1996) Diversity among populations of switchgrass based on RAPD markers. Crop Sci 36:1017–1022

Hultquist S, Vogel KP, Lee D, Arumuganathan K, Kaeppler S (1996) Chloroplast DNA and nuclear DNA content variations among cultivars of switchgrass, *Panicum virgatum* L. Crop Sci 36:1049–1052

Jurka J, Pethiyagoda C (1995) Simple repetitive DNA sequences from primates: compilation and analysis. J Mol Evol 40:120–126

Kantety R, La Rota M, Matthews D, Sorrells M (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48:501–510

Kellogg E (2001) Evolutionary history of the grasses. Plant Physiol 125:1198–1205

Lazo GR et al (2004) Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. Genetics 168:585–593

Lerner DR, Raikhel NV (1989) Cloning and characterization of root-specific barley lectin. Plant Physiol 91:124–129

McLaughlin SB, Walsh ME (1998) Evaluating environmental consequences of producing herbaceous crops for bioenergy. Biomass Bioenergy 14:317–324

Moore KJ, Moser LE, Vogel KP, Waller SS, Johnson BE, Pedersen JF (1991) Describing and quantifying growth stages of perennial forage grasses. Agron J 83:1073–1077

Moser LE, Vogel KP (1995) Switchgrass, big bluestem, and indiangrass. In: Barnes RF, Miller DA, Nelson CJ (eds) An introduction to grassland agriculture, chapter 32. Iowa State University Press, Ames, pp 409–420

Picoult-Newberg L, Ideker T, Pohl M, Taylor S, Donaldson M, Nickerson D, Boyce-Jacino M (1999) Mining SNPs from EST databases. Genome Res 9:167–174

Richards HA, Rudas VA, Sun H, McDaniel JK, Tomaszewski Z, Conger BV (2001) Construction of a GFP-BAR plasmid and its use for switchgrass transformation. Plant Cell Rep 20:48–54

Saha M, Mian M, Eujayl I, Zwonitzer J, Wang L, May G (2004) Tall fescue EST-SSR markers with transferability across several grass species. Theor Appl Genet 109:783–791

Somleva MN, Tomaszewski Z, Conger BV (2002) *Agrobacterium*-mediated genetic transformation of switchgrass. Crop Sci 42:2080–2087

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441–1452

Vogel KP, Haskins FA, Gorz HJ, Anderson BA, Ward JK (1991) Registration of "Trailblazer" switchgrass. Crop Sci 31:1388

Vogel KP, Hopkins AA, Moore KJ, Johnson KD, Carlson IT (1996) Registration of "Shawnee" switchgrass. Crop Sci 36:1713

Weber J (1990) Informativeness of human (dC-dA)*n*.(dG-dT)*n* polymorphisms. Genomics 7:524–530

Wu X-L, Griffin K, Garcia M, Michal J, Xiao Q, Wright R, Jiang Z (2004) Census of orthologous genes and self-organizing maps of biologically relevant transcriptional patterns in chickens (*Gallus gallus*). Gene 340:213–225

Yu J, La Rota M, Kantety R, Sorrells M (2004) EST derived SSR markers for comparative mapping in wheat and rice. Mol Genet Genomics 271:742–751